INCORPORATING LANGUAGE LEVEL INFORMATION INTO ACOUSTIC MODELS

Peidong Wang, Deliang Wang

Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210

ABSTRACT

This paper proposed a class of novel Deep Recurrent Neural Networks which can incorporate language-level information into acoustic models. For simplicity, we named these networks Recurrent Deep Language Networks (RDLNs). Multiple variants of RDLNs were considered, including two kinds of context information, two methods to process the context, and two methods to incorporate the language-level information. RDLNs provided possible methods to fine-tune the whole Automatic Speech Recognition (ASR) system in the acoustic modeling process.

Index Terms— RDLN, ASR, HMM, Viterbi Decoding

1. INTRODUCTION

The past few years have witnessed the successful application of Deep Neural Networks (DNNs) to Automatic Speech Recognition (ASR) tasks [1]. The conventional DNN based ASR system has a Hidden Markov Model (HMM) [2] to deal with the variant temporal transitions, forming a DNN-HMM hybrid model.

Although various Recurrent Neural Networks (RNNs) based alternatives to DNN-HMM hybrid models have been proposed [3], as far as we know, none of them prevailed the generalized DNN-HMM hybrid models. The generalized DNN-HMM hybrid models include all models replacing DNN with other frameworks like Recurrent Neural Networks (RNNs) [4], Convolutional Neural Networks (CNNs) [5], and Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNNs) [6].

The way we train the HMM in the conventional DNN-HMM hybrid model is to first use a Gaussian Mixture Model (GMM) as the acoustic model. We can get the force-aligned phoneme-level labels during this process. After that, we train the DNN acoustic model using the frame-wise inputs and the force-aligned labels. At the test stage, we pass the features extracted from speech frames and get the probability of the frame being different phonemes from the output of the DNN model. After that, the outputs of the DNN on different frames are fed into a Viterbi decoder to get the words.

Note that in the process above, the HMM model does not evolve in both DNN training and test stages. In addition, there is a mismatch between the training objective, which is to minimize the differences between DNN outputs and the phoneme-level labels, and the model evaluation criteria, which is to reduce the Word Error Rate (WER).

We view both Connectionist Temporal Classification (CTC) [7] and RNN Encoder-Decoders [3] as beneficial attempts to solve the second problem above. And in this paper, we propose an alternative method using Recurrent Deep Language Networks (RDLNs). We will explain the idea and variants of RDLNs in the next section. Then we will show the experimental results and make a conclusion.

2. INCORPORATING LANGUAGE LEVEL INFORMATION INTO ACOUSTIC MODELS

2.1. Language-Level Information

As mentioned in the previous section, a long-time problem in the DNN-HMM based ASR system is the mismatch between training objectives and evaluation criteria for the DNN acoustic model. So in order to solve this problem, a natural question is "Where is the language-level information?". We will quickly find that we use language-level information only in the Viterbi decoding process at the test stage.

In Viterbi decoding process, we have the following equation to calculate the probability of HMM state j at frame i.

$$P_{i,j} = (\sum_k P_{i-1,k} * T_{k,j}) * O_{i,j} \tag{1}$$

The $T_{k,j}$ in the above equation denotes the transition probability from HMM state k to HMM state j, which may vary according to different frames in a context-dependent decoder like Kaldi [8]. The $O_{i,j}$ corresponds to the output of DNN. Viterbi decoder uses a backward propagation process using the probabilities calculated in (1), to find the path that has the largest probability.

If we set all values of $O_{i,j}$ in (1) to be the same, for example 1, we will notice that the remaining part in the equation can be viewed as a prediction of the probabilities of HMM states at frame i. This indicates that we can use the decoder in a different way and get the language-level information we want.

$$P_{i,j}^{pred} = (\sum_{k} P_{i-1,k} * T_{k,j}) * (1/z) \quad (2)$$

z in (2) is a normalization value, which is the same for all j's.

After getting the language-level information, we can choose to build the RDLN model from at least two kinds of context information, two methods to process the context, and two methods to incorporate the language-level information.

2.2. Context Information Selection

The context information in RDLN model corresponds to how do we get the $P_{i-1,k}$'s in (2). We can either use the DNN outputs of the previous frames, or use the labels of the previous frames. Using the labels will usually reduce the computing complexity and make the additional information purely from the language-level. But the advantage of using the real outputs is that it will incorporate a phoneme-level information in addition to language-level. We may choose one kind of context information according to our needs.

2.3. Process the Context

We can use either context-dependent or context-independent methods to process the context information in the previous section.

To process the context information in a context-dependent way, we may simply use the Viterbi decoder and set the outputs of current frame to be a same value, like 1. Then we take the HMM state probability generated by the decoder as the processed context. Note that after the decoding process, the processed context now to some extend contains the language-level information.

Another way to process the context information is in a context-independent method. We can get the transition matrix T and use it to transform the context information obtained in the previous section. This method is comparable to the first method, as indicated in the Kaldi documentations.

2.4. Incorporate Language-Level Information

There are at least two ways to incorporate the language-level information into acoustic models.

The first way is through a modified objective function. We may use the labels based language-level information as an additional target, and real outputs based language-level information as an additional estimation to the target.

The second way is to stack the information into the input vectors. In this way, the outputs may be more accurately and robustly estimated by the DNN model.

3. EXPERIMENTS

We conducted experiments on all single channel utterances of CHiME-4 dataset. We used the real outputs as the context information, processed the context information in a context-independent manner, and incorporate the information as additional inputs to the DNN model. In the experiments, I only used the context information of one frame previous to the current one for simplicity, but RDLNs can take on as much context information as we need.

The way we obtained the transition matrix T in Kaldi is as follows. The outputs of DNN are denoted as pdf-id's. And since we only used one previous frame, we have $P_{0,k} = O_{0,k}$. So that our problem is to find the transition matrix from $O_{0,k}$ to $P_{1,j}$. Since there is no relation between pdf-id's and the triphone states, we need to first convert pdf-id's to the transition-id's in Kaldi. Then we use the transition relation represented in transition-id's to transform the state. After that, we convert back from transition-id's to pdf-id's.

Note that after the steps above, the additional input had a dimension of 3161. So we compressed it into a 42 dimensional vector using the relationship between pdf-id's and monophone states. We used the 8th epoch of the baseline model as the initial model to RDLN.

Some preliminary results are shown in Figure 1.

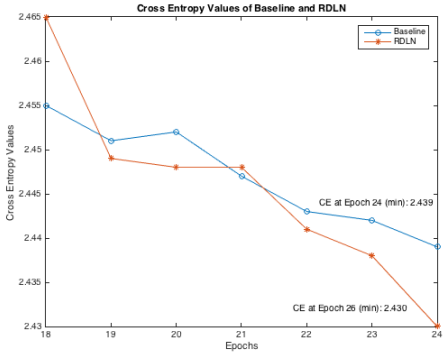

Fig. 1: Cross Entropy Values of Baseline and RDLN

From Figure 1, we can see that the Cross Entropy values of RDLN is constantly lower than the baseline system after epoch 19, with only one out-lier. In addition, the improvement gets larger after each training epoch. Note that in the experiment above, we only incorporated one previous frame. It is of high possibility that we will get better results using more context information.

4. CONCLUSION

This paper proposed a method to incorporate language-level information into acoustic models for a DNN-HMM hybrid ASR system. We can fin-tune the whole ASR system, instead of only the acoustic modeling part, using our method.

Then we discussed about two kinds of context information, two methods to process the context, and two methods to incorporate the language-level information. Experiments showed that adding language-level information into acoustic models constantly improved the performance over the baseline system. We can foresee a lot of possible contributions along this technical path, including the comparison of the eight variants of RDLN and the substitution of the conventional DNN by RNNs and CNNs. We also want to extend this method so that it can be used in non-HMM based frameworks.

5. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Sean R Eddy, "Hidden markov models," *Current opinion in structural biology*, vol. 6, no. 3, pp. 361–365, 1996.

[3] Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: first results," *arXiv preprint arXiv:1412.1602*, 2014.

[4] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 6645–6649.

[5] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, and Gerald Penn, "Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition," in *2012 IEEE international conference on Acoustics, speech and signal processing (ICASSP)*. IEEE, 2012, pp. 4277–4280.

[6] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed, "Hybrid speech recognition with deep bidirectional lstm," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 273–278.

[7] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[8] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.